Programme des enseignements

3e année

Filière Statistique pour les sciences de la vie

ANNÉE SCOLAIRE 2019 / 2020

ENSAI

École nationale
de la statistique
et de l'analyse
de l'information

FILIÈRE STATISTIQUE POUR LES SCIENCES DE LA VIE

ANNÉE SCOLAIRE 2019/2020

BIOSTATISTICS SPECIALIZATION

2019/2020 ACADEMIC YEAR

Table des matières

Présentation de la filière

Thématiques abordées

Pour répondre aux exigences de la filière, en plus du tronc commun à tous les élèves de troisième année, l'enseignement est divisé en six unités d'enseignement (UE).

Après des compléments en statistique, notamment en données de survie, modèles mixtes, régularisation et plans d'expériences, les enseignements apportent les outils nécessaires pour une spécialisation dans le domaine de l'expérimentation. Les cours d'épidémiologie, d'essais cliniques et l'analyse des données Omics permettent en particulier aux étudiants de recevoir une solide formation pour des applications dans le secteur de la santé.

Tout au long de l'année, les étudiants auront à gérer différents projets. Les trois principaux correspondent aux cours d'épidémiologie, d'essais cliniques et l'analyse des données Omics. Ils permettent de compléter et de mettre en pratique les connaissances théoriques vues en cours. Ces projets sont aussi l'occasion de développer l'esprit d'équipe par des travaux en groupes de plusieurs étudiants.

Transversalement à ces unités d'enseignement, les applications en informatique (R, SAS, WINBUGS…etc) sont omniprésentes. Des séminaires professionnels présentent la richesse des métiers offerts en statistique pour les sciences de la vie. Ils sont l'occasion d'une présentation, par les praticiens, des outils ou modèles statistiques utilisés dans les entreprises et instituts de recherche.

La langue anglaise n'est pas négligée puisque des enseignements donnent lieu à l'écriture de mémoires en anglais et deux enseignements sont dispensés dans cette langue.

Option Formation Par la Recherche

L'Ensai offre la possibilité, aux élèves de 3ème année qui le souhaitent, de se préparer en vue d'une carrière de chercheur en entreprise au sein des services Recherche et Développement ou dans le secteur académique. Dans le cadre de l'option formation par la recherche (Ofpr), ces élèves bénéficient d'un aménagement de leur scolarité leur permettant de suivre au choix un des trois masters recherche suivants :

- Master Recherche, mention Mathématiques et applications, spécialité Statistique Mathématique
- Master Recherche, mention Santé Publique, spécialité Modélisation en Pharmacologie Clinique et Epidémiologie (MPCE)
- Master Recherche, mention Bio-informatique, spécialité Bio-informatique en santé

A l'issue de ce parcours, ils pourront poursuivre cette formation par une thèse académique ou de type Cifre (Convention Industrielle de Formation par la Recherche).
Les thèses académiques sont en général encadrées dans des laboratoires de recherche tels que ceux du CNRS ou de l'Inserm. En ce qui concerne les entreprises signataires de thèses Cifre ou organismes de recherche, on peut citer par exemple l'I.R.I.S. –Laboratoires Servier associé à l'INSERM ou encore l'IRSN associé à l'Université de Paris 11.

Les entreprises partenaires

La filière bénéficie de partenariats avec des acteurs économiques de premier plan. Ces partenariats permettent de développer des échanges privilégiés notamment via des cours, des séminaires professionnels et des stages.


QUINTILES
Keyrus
BIOPHARMA
SANOFI
SERVIER
SOLADIS

Volume horaire						
	Cours	Ateliers	Projets	Total	Crédits	Enseignant-e-s
UE0 Tronc commun						
Droit du Travail	3	6		9	0.5	Charlotte GRUNDMAN
Anglais	30			30	1.5	
Sport		30		30	0	
Total	33	36	0	69	2	
UE1 Machine learning						
Machine learning	18	21		39	3	Hong-Phuong DANG, Romaric GAUDEL, Fabien NAVARRO, Brigitte GELEIN
Régression pénalisée et sélection de modèles	9	6		15	1	
Apprentissage statistique à grande échelle	9	9		18	1.5	Arthur KATOSSKY, Rémi PÉPIN
Webmining et traitement du langage	9	12		21	1.5	Arthur KATOSSKY
Total	45	48	0	93	7	
UE2 Méthodologie statistique 1						
Plans d'expériences	18			18	1	Walter TINSSON
Modèles mixtes	15	9		24	1.5	Etienne DANTAN
Modèles de durée	18	6		24	1.5	Lisa BELIN
Total	51	15	0	66	4	
UE3 Méthodologie statistique 2						
Mesures de qualité de vie	12	6		18	1	Jean-Benoit HARDOUIN
Statistique des processus	12			12	1	Myriam VIMOND
Statistique des processus — Compléments (SV)		3		3	0	Myriam VIMOND
Traitement des données manquantes	6	6		12	1	Mélanie PRAGUE
Statistique bayésienne	9	6		15	1	Sophie ANCELET, Eric PARENT
Méta-analyse	12	6		18	1	Drifa BELHADI
Total	51	27	0	78	5	
UE4 Essais cliniques						
Essais cliniques	18			18	1	Damien-David HAJAGE, Lisa BELIN ; Yann DE RYCKE
Pharmacométrie	6	12		18	1	Jérémie GUEDJ, Thu Thuy NGUYEN
Projet d'essais cliniques		6	18	24	2	Yann DE RYCKE, Damien-David HAJAGE
Total	24	18	18	60	4	
UE5 Épidémiologie						
Épidémiologie quantitative	18			18	1	Olivier GRIMAUD, Pascal CREPEY, Cindy PADILLA
Modélisation compartimentale	6	6		12	1	Audrey LAVENU
Total	24	6	0	30	2	
UE6 Statistique pour données Omics						
Analyse de données -omics	6	12		18	1	Isabel BRITO, Pierre NEUVIAL
Introduction à l'analyse des données Omics	6	6		12	1	Guillaume DESACHY
Total	12	18	0	30	2	
Projet de fin d'étude						
Projet méthodologique		3	9	12	1	
Projet de fin d'étude		9	27	36	3	
Data Challenge		12		12	0	
Total	0	24	36	60	4	
Séminaire professionnel						
Séminaire professionnel	30			30	0	
Total	30	0	0	30	0	
TOTAL	270	192	54	516	30	

UE Stage 3A	Crédits
	25

Descriptifs des enseignements du tronc commun

UE0 – Tronc commun

Droit du Travail

Work Law

Cours : 3h • Atelier : 6h

Enseignant : Charlotte GRUNDMAN, Avocat au Barreau de Paris.

Correspondant : Ronan Le Saout

Objectif pédagogique

La matière étant extrêmement vaste et complexe, il est ici proposé aux étudiants une approche didactique et vivante du sujet, l'objectif de l'enseignement étant de permettre aux étudiants qui travailleront dans un futur proche en entreprise d'avoir compris certaines notions pratiques essentielles en droit du travail.

Contenu de la matière

A cette fin, et hormis le cours d'amphi, il sera systématiquement proposé aux étudiants, après l'étude d'une notion, un exercice visant à mettre en pratique la notion abordée.

Afin de satisfaire le plus possible à cet objectif, il est ainsi proposé l'organisation suivante des cours :

Cours commun (3 heures) :

Chapitre 1 : Comprendre d'où l'on vient pour savoir où on va :

- Introduction historique au droit du travail

- Les sources du droit du travail
  - sources imposées,
  - sources négociées
- Ordre public absolu et ordre public social

Chapitre 2 : les instances de contrôle du droit du travail

- L'inspecteur du travail
- Les multiples juges du droit du travail
- Point sur la procédure prud'homale

Chapitre 3 : Formation et exécution du contrat de travail

- la qualification du contrat de travail : « faux artisans, faux auto-entrepreneurs et vrai salarié ».
- le contrat à durée indéterminée, norme juridique et sociale
- la période d'essai après la loi du 25 juin 2008 : définition, durée et rupture
- les principales clauses du contrat de travail :
- la clause de mobilité
- la clause de non-concurrence

Chapitre 4 : la rupture du contrat à durée indéterminée

- le licenciement pour motif personnel
- le licenciement pour motif économique
- la démission du salarié
- les autres modes de rupture

Les TD :

La première heure de cours sera consacrée à l'étude d'un chapitre. Cet exposé sera suivi d'une mise en situation pratique, où les étudiants devront par groupe répondre à un cas pratique. Un rapporteur sera désigné par groupe, et la notation se fera à cette occasion.

Chapitre 1 : La modification du contrat de travail

Modification du contrat de travail et changement des conditions de travail

- la durée du travail (focus sur le forfait-jour)
- le lieu de travail
- la rémunération

Chapitre 2 : Le recrutement

Chapitre 3 : les droits fondamentaux du salarié

- Le fait religieux en entreprise
- Vie personnelle et technologies de l'information et de la communication (TIC)
- La mise en place de moyens de contrôle via les TIC en entreprise
- Harcèlements
- Maladie et maternité du salarié

Langue d'enseignement
Français

UE0 – Tronc commun

Anglais

English

Cours : 30h (dont 15h d'aide au projet)

Enseignant	:	Divers intervenants
Correspondant	:	Todd DONAHUE

Objectif pédagogique

Les élèves qui n'ont pas passé ou qui n'ont pas réussi le TOEIC l'année dernière auront progressé dans les compétences requises – c'est à dire, la compréhension orale, la reconnaissance des erreurs, les pièges grammaticaux, et la compréhension écrite. Les autres auront acquis les compétences nécessaires pour affronter le monde professionnel. Ils auront vu les tournures qui aident à diriger et à participer à des réunions, à prendre des décisions, et à négocier. Ils se seront entraînés à faire des présentations. Ils auront rédigé un projet en anglais et préparé la soutenance de ce projet.

Contenu de la matière

Pour les élèves qui n'ont pas eu un score d'au moins 785 au TOEIC : pendant les 5 premières séances, la plupart des cours seront basés sur la préparation à cet examen. Les ressources informatiques de l'Ecole doivent aussi être mises à profit (pages Moodle, TOEIC Mastery), ainsi que les méthodes disponibles à la bibliothèque.

Pour les autres élèves, les cours seront organisés par groupe de niveau et conçus afin de les préparer à affronter le monde professionnel sur le plan international. Les thèmes suivants seront traités : « Leading meetings », « Interviews », « Presentations », « Taking decisions », et « Negociating deals », et « Cultural and Political Current Events ».

Ensuite, les 5 dernières séances seront consacrées au travail de rédaction/correction des rapports faits en anglais dans chaque filière ainsi qu'à la préparation des soutenances orales. Chaque responsable de filière indiquera aux élèves, en début d'année, le projet concerné et les modalités de notation. Les élèves recevront des consignes détaillées avant de démarrer ces cinq séances, afin d'arriver à la première séance avec une première version ou extrait de leur rapport en anglais prêt pour correction et relecture.

Pré-requis

Aucun

Contrôle des connaissances

L'examen final prend la forme d'une simulation d'entretien d'embauche. Cet examen oral durera environ 25 minutes, sera noté, et permettra d'évaluer le niveau d'expression orale sur l'échelle CECRL*. Le CV et la lettre faite pour cet exercice seront évalués et feront partie de la note finale. Le niveau acquis apparaîtra sur le Supplément au diplôme. L'objectif de la CTI† pour tous les élèves ingénieurs est d'atteindre le niveau B2.

* le **C**adre **e**uropéen **c**ommun de **r**éférence pour les **l**angues.
† la **C**ommission des **T**itres d'**I**ngénieur.

Références bibliographiques

- Arbogast, B., *30 Days to the TOEIC Test*, Canada: Peterson's, 2002.
- Schrampfer-Azar, B., *Understanding and Using English Grammar*, New York: Longman, 1999.
- Buckwalter, Elvis, et.al, *Boostez votre score au TOEIC-spécial étudiants*, Paris: Eyrolles, 2009.
- Gear, Jolene, *Cambridge Grammar and Vocabulary for the TOEIC Test*, Cambridge: Cambridge University Press, 2010.
- Lecomte, Stéphane, et. al, *La Grammaire au TOEIC et au TOEFL : Mode d'emploi*, Paris: Ophrys, 2008.
- Lougheed, Lin, *Tests complets pour le nouveau TOEIC* (4ème ed.), Paris: Pearson Education France, 2008.
- MBA Center, *New TOEIC Study Book*, Paris: MBA Center Publications, 2007.

Langue d'enseignement

Anglais

Pour tout complément d'information, chaque élève peut consulter le Programme des enseignements : Langues étrangères, distribué au début de l'année académique.

UE0 – Tronc commun

Sport

Sport

TP : 30h

Enseignant : Divers intervenants

Correspondant : Julien LEPAGE

Cours facultatif

Objectif pédagogique

L'objectif est d'amener les élèves à maintenir un esprit sportif, sortir du strict cadre académique et développer leurs capacités physiques.
Contenu de la matière

9 activités sportives sont proposées par l'école :

- Badminton
- Basket
- Cross-Training
- Football
- Hand-ball
- Tennis de table
- Tennis débutant
- Volley-ball
- Course à pied/préparation physique/coaching sportif

Outre les entraînements, les élèves inscrits peuvent être amenés à participer à des compétitions.

Prise en compte dans la scolarité
La participation à une activité sportive peut donner lieu à l'attribution d'un bonus ajouté sur la moyenne du semestre concerné. Le niveau de ce bonus est précisé dans une circulaire d'application en début d'année académique. Il varie selon l'assiduité aux séances, l'engagement et la participation aux compétitions tout au long de l'année.
Pour être définitive, la liste des élèves bénéficiant de ces bonus doit être validée par le directeur des études.

Un bonus peut être exceptionnellement attribué en dehors des activités sportives réalisées dans le cadre Ensai. Pour y prétendre, les élèves concernés doivent remplir les 3 conditions suivantes :
- pratiquer régulièrement une activité sportive et participer aux compétitions liées ;
- posséder un niveau national (voir très bon niveau régional suivant le sport en question) ;
- déposer une demande argumentée auprès de la direction des études et du service sport en début d'année scolaire, afin de faire valider le programme d'entraînement, des compétitions et les modalités de diffusion des performances. Pour certains ayant des contraintes sportives, des aménagements horaires pourront d'ailleurs être ainsi envisagés si besoin.

UE1 – Machine Learning

Machine Learning

Machine Learning

Cours : 18h - Atelier : 21h

Enseignants	:	Hong-Phuong DANG (Ensai), Romaric GAUDEL (Ensai), Fabien NAVARRO (Ensai) et Brigitte GELEIN (Ensai)
Correspondant	:	Arthur KATOSSKY (Ensai)

Objectif pédagogique

Ce cours présente les principes de l'apprentissage automatique (Machine Learning) ainsi que les modèles les plus utilisés.

Contenu de la matière

- Principes de l'apprentissage automatique
  - Apprentissage supervisé vs. non-supervisé ; échantillon d'entraînement et de validation, overfitting, erreur de généralisation ; fonction de coût (loss function) et minimisation d'une erreur ; évaluation des méthodes non-supervisées ; méthodes vues en 2A en tant que méthodes d'apprentissage
- Réseaux de neurones
  - Principe des réseaux de neurones ; propriétés des réseaux de neurones simples ; descente de gradient ; réseaux de neurones profonds ; architectures particulières (ex: réseaux de convolution ; réseaux récurrents ; ...) ; réduction de la dimension à l'aide de réseaux de neurones (auto-encodeurs ; word2vec ; ...).
- Méthodes d'agrégation
  - Quelques rappels et approfondissements (CART, multiregression trees), Bagging, random forests, Boosting, XGBoost, Stacking (agrégation de modèles de types différents par construction d'un modèle « superviseur » qui combine au mieux les prédictions des modèles primaires.)
- Support Vector Machines
  - Classification par hyper-plan séparateur ; classifieur de marge maximale ; données non linéairement séparable et méthodes à noyau ; SVM multi-classe ; liens avec d'autres modèles (logistique, réseaux de neurones) ; descente de gradient

Compétences

- Identifier comment résoudre une tâche par apprentissage automatique
- Choisir un modèle a priori adapté à une tâche
- Utiliser un modèle de l'état de l'art (SVM, réseau de neurones, forêt, …)
- Comparer empiriquement différents modèles pour une tâche donnée

Pré-requis

R, Python, algèbre linéaire, optimisation de fonctions

Contrôle des connaissances

Des TP notés + un examen final

Références bibliographiques

- Endrew Ng. Machine Learning Yearning. Disponible gratuitement au lien https://www.deeplearning.ai/machine-learning-yearning/.
- Rémi Gilleron. Apprentissage machine - Clé de l'intelligence artificielle - Une introduction pour non-spécialistes. Ellipses.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. 2016

Langue d'enseignement

Français

UE1 – Machine Learning

Régression pénalisée et sélection de modèles

Penalized problems and model selection

Cours : 9h - Atelier : 6h

Enseignants : Cédric HERZET (INRIA) & Clément ELVIRA (INRIA)

Correspondant : Arthur KATOSSKY (Ensai)

Objectif pédagogique

De nombreuses tâches d'apprentissage et de traitement du signal visent à retrouver un ensemble de grandeurs inconnues (état d'un système, modèle génératif, etc) à partir de données.
Malheureusement, dans de nombreuses situations, les données disponibles s'avèrent insuffisantes pour lever l'ambiguïté sur les quantités à inférer ou les estimer avec une précision suffisante.
Une manière classique de contourner ce problème consiste à introduire une information « a priori » sur la solution recherchée.
Plus particulièrement, dans ce cours nous montrons comment lever l'ambiguïté inhérente à certains problèmes en « pénalisant » les solutions ne présentant pas certaines caractéristiques d'intérêt.
L'objectif de ce cours est d'identifier et manipuler les méthodes de pénalisation les plus courantes.

Contenu de la matière
- Identifier la pénalisation la plus adaptée à une tâche
- Résoudre un problème d'optimisation comportant un terme de régularisation
- Régler les paramètres du modèle

Pré-requis
- Algèbre linéaire
- Optimisation convexe
- Programmation en Python

Contrôle des connaissances
TP notés + examen final

Références bibliographiques
- C. Bishop. Pattern recognition and machine learning. Springer-Verlag New York, 2006.
- S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing. Applied and Numerical Harmonic Analysis. Birkhau¨ser, 2013.
- D. P. Bertsekas. Nonlinear Programming. Athena Scientific, USA, 2003.

Langue d'enseignement

Français

UE1 – Machine Learning

Apprentissage statistique à grande échelle

Large-scale Machine-Learning

Cours : 9h - Atelier : 9h

Enseignant : Arthur KATOSSKY (Ensai) et Rémi PEPIN (Ensai)

Correspondant : Arthur KATOSSKY (Ensai)

Objectif pédagogique

Au cours de la dernière décennie, nous avons assisté à l'émergence d'applications numériques nécessitant de faire face à de gigantesques quantités de données, générées de plus en plus rapidement. Ces applications (surveillance de réseaux, biologie et médecine, applications financières, réseaux sociaux, etc.) nécessitent un besoin grandissant de techniques capables d'analyser et de traiter ces grandes masses d'information, avec précision et efficacité. La statistique rejoint ici les sciences du numérique, et plus précisément l'informatique répartie, pour proposer de nouvelles approches, relatives au Big Data. Les techniques et les modèles doivent prendre en compte le volume pléthorique de ces données, mais également leur génération rapide en continu (vélocité) ainsi que la diversité de leur format (variété) et la qualité de l'information (véracité), appelés communément les 4V du Big Data.

Contenu de la matière

- Les différents « v »
- Principes, avantages et inconvénients d'un système réparti
- connaître les stratégies de tolérance aux fautes (duplication des données, exécution avec erreurs)

Compétences

- Identifier l'architecture adaptée à une tâche (exécution séquentielle et/ou parallèle, exécution en mémoire et/ou en flux, exécution locale et/ou distante).
- Lancer des calculs sur une architecture Big Data (notamment, appliquer les paradigme Map-Reduce).
- exécuter des calculs volumineux – et en particulier des calculs statistiques – sur des prestataires de calcul (IaaS ou PaaS comme Amazon Web Service, Google Cloud Platform ou autre)

Pré-requis

Algorithmique.

Contrôle des connaissances

À déterminer.

Références bibliographiques

- Analyses des Big Data : quels usages, quels défis ? Note d'analyse du Comissariat général à la stratégie et la prospective
- Pirmin Lemberger, Marc Batty, Médéric Morel, Jean-Luc Raffaëlli. Big Data et machine learning - Manuel du data scientist, Dunod, 2015.
- Rudi Bruchez. Les bases de données NoSQL et le BigData : Comprendre et mettre en œuvre, Eyrolles (2015)

Langue d'enseignement

Français.

UE1 – Machine Learning

Webmining et traitement du langage

Webmining & NLP

Cours : 9h - Atelier : 12h

Enseignant	:	Arthur KATOSSKY (Ensai)
Correspondant	:	Arthur KATOSSKY (Ensai)

Objectif pédagogique

Le cours de *webmining & natural language processing* répond à plusieurs objectif:
- pratiquer la collecte de données, l'extraction d'information et l'appariement de sources
- équiper les élèves avec des outils théoriques pour l'étude des données textuelles
- faire comprendre les grandes approches qui structurent le foisonnement de modèles de la langue
- présenter des exemples concrets d'applications dans les différentes domaines d'application des élèves
- donner la capacité de réaliser des tâches classiques en étude de texte: classification, analyse de sentiment, détection d'entités, etc.

Contenu de la matière
- Introduction au traitement automatique du langage (*natural language processing*)
- Grandes catégories de modèles : *bag-of-words* et *tf-ifd* ; réseaux de neurones (LSTM, GRU, etc.) ; plongements de mots (word2vec, GLoVe, fasttext, Elmo, BERT, etc.) ; modèles probabilistes (HMM, CRF, LDA, etc.)
- Applications : classification, analyse de sentiment, détection d'entités, etc.
- Traitement de donnés textuelles et extraction d'information
- Collecte de données sur le web et utilisation d'une API

Pré-requis

Apprentissage statistique (réseaux de neurones) ; apprentissage statistique à grande échelle ; statistique bayésienne ; chaînes de Markov

Contrôle des connaissances

Projet

Références bibliographiques

Communiquée ultérieurement

Langue d'enseignement

Français.

Descriptifs des enseignements de la filière

UE2 - Méthodologie statistique 1

Plans d'expériences

Experiment Design

Cours : 18h

Enseignant	Walter TINSSON (Université de PAU)
Correspondant	Lionel TRUQUET

Enseignement destiné aux élèves des filières « Statistique pour les sciences de la vie» et « Génie statistique » , et « Ingénierie statistique des territoires et de la santé »

Objectif pédagogique

Comprendre les principes fondateurs des stratégies d'expérimentation
Apprendre à choisir, construire un dispositif expérimental
Acquérir les outils d'analyse des plans d'expériences (utilisation du logiciel R)

Contenu de la matière

Principes fondateurs et présentation des grandes familles de plans
Les Outils d'Analyse : modèle linéaire, analyse de la variance
Plans factoriels complets et fractionnaires, Optimalité
Expériences Accélérées

Pré-requis

Ce qui a été vu en classes préparatoires et lors des deux premières années de l'Ecole est largement suffisant pour suivre le cours (Calcul matriciel, Optimisation, Bases de la régression et d'analyse de variance…)

Contrôle des connaissances

Examen pratique en salle informatique

Références bibliographiques

- AZAIS, J.-M., BARDET, J.-M. Le modèle linéaire par l'exemple (2e éd.). Dunod, 2012.
- DROESBEKE, J.-J., FINE, J., SAPORTA, G. (Eds Scientifiques). Plans d'expériences : Applications à l'entreprise. Technip, 1997.

Langue d'enseignement

Français

UE - Méthodologie statistique 1

Modèles mixtes

Mixed Models

Cours : 15h • Atelier : 9h

Enseignant : Etienne DANTAN (Université de Nantes)
Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

A l'issue de cet enseignement, les élèves devront connaître les fondements de la théorie statistique du modèle mixte afin d'en assurer une bonne compréhension, maîtrise et interprétation.

Contenu de la matière

Dans le modèle linéaire, on peut prendre en compte diverses structures de corrélation (intra-classe, temporelle, spatiale) grâce à une version dite "mixte" du modèle qui fait intervenir à la fois des effets fixes et des effets aléatoires.

Il est à noter que nombre de problèmes peuvent être abordés sous une formulation de modèle mixte, cf. le filtre de Kalman et les splines cubiques par exemple. Ce type de modèles peut faire l'objet d'une grande variété de traitements et d'interprétation statistique (algorithme EM, méthodes MCMC, approche bayésienne). Le concept permet également de nombreuses extensions (modèle linéaire généralisé, modèle non linéaire).

Le modèle mixte connaît actuellement un développement important de ses applications dans maints secteurs de l'industrie, des sciences économiques et sociales, de la biologie et de la médecine. Il est servi par de bons logiciels tels que les Proc GLM, Mixed, Nlmixed, Glimmix de SAS, Asreml , Nlme de R et S plus, Monolix, Genstat et Winbugs.

1. Introduction: écriture générale, hypothèses, exemples
2. Prédiction des effets aléatoires: meilleure prédiction, meilleure prédiction linéaire et meilleure prédiction linéaire sans biais (BLUP)
3. Equations du modèle mixte d'Henderson
4. Inférence par la méthode du maximum de vraisemblance : estimation et tests d'hypothèse des effets fixes et des composantes de variance
5. Concept de vraisemblance résiduelle (REML) : présentation classique et présentation bayésienne
6. Théorie de l'algorithme EM
7. Application aux composantes de variance
8. Aperçu sur le modèle linéaire généralisé mixte

Pré-requis

Lois de probabilité, algèbre linéaire et théorie du modèle linéaire (regréssion, anova)

Contrôle des connaissances

Examen écrit.

Références bibliographiques

- B.P CARLIN, T.A LOUIS, *Bayesian Methods for Data Analysis*, Chapman & Hall/CRC Press, 3rd edition, 2009
- A.P. DEMPSTER, N.M. LAIRD, D.B. RUBIN*, Maximum likelihood from incomplete data via the EM algorithm*, J. R. Statist. Soc. B 39 1-38, 1977
- P.J DIGGLE, P HEAGERTY, K-Y LIANG, S ZEGER, *Analysis of Longitudinal Data*, Oxford Statistical Science Series, 2nd edition, 2002
- L. FAHRMEIR, G. TUTZ, Multivariate statistical modelling based on generalized linear models (2nd ed.), Springer Verlag, Berlin, 2001
- H.O. HARTLEY, J.N.K. RAO, Maximum likelihood estimation for the mixed analysis of variance model, Biometrika 54 93-108, 1967
- M. FITZMAURICE, N LAIRD, J.H. WARE, *Applied Longitudinal Analysis (2nd ed.)*, Wiley series in probability and statistics, 2011
- D.A. HARVILLE, Maximum likelihood approaches to variance component estimation and related problems, J. Am. Stat. Assoc. 72, 320-338, 1977
- C.R. HENDERSON, *Applications of linear models in animal breeding,* University of Guelph, Guelph, 1984 (en ligne)
- N.M. LAIRD, J.H. WARE, *Random effects models for longitudinal data*, Biometrics 38 963-974, 1982
- K.Y. LIANG, S.L. ZEGER, Longitudinal data analysis using generalized linear models, Biometrika 73 13-22, 1986
- P. Mc CULLAGH, J.A. NELDER, *Generalized linear models*, 2nd edition, Chapman & Hall, London, 1989
- G. MOLENBERGHS, G. VERBEKE, *Models for Discrete Longitudinal Data*, Springer Verlag, New York, 2005

- H.D. PATTERSON, R. THOMPSON, Recovery of interblock information when block sizes are unequal, Biometrika 58 545-554, 1971
- J.C. PINHEIRO, D.M. BATES, *Mixed effects models in S and S-plus*, Springer, Berlin, 2000
- C.R. RAO, J. KLEFFE, *Estimation of variance components and applications,* North Holland Series in Statistics and probability, Elsevier, 1988
- C. ROBERT, Méthodes de Monte Carlo par chaînes de Markov, Economica, Paris, 1996
- S.R. SEARLE, G. CASELLA, C.E. Mc CULLOCH, *Variance components,* J Wiley & Sons, New York, 1992
- G. VERBEKE, G. MOLENBERGHS, *Linear mixed models in practice*, Springer Verlag, New York, 1997

Langue d'enseignement
Français

UE2 - Méthodologie statistique 1

Compléments de modèles de durée

Survival Analysis Applied to Biostatistics

Cours : 18h • Atelier : 6h

Enseignant : Lisa BELIN (INSERM Hôpital Bichat)

Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

A l'issue de cet enseignement, les élèves devront maîtriser les connaissances de base pour analyser des données censurées. Ces problèmes concernent tous ceux qui sont impliqués dans les sciences de la vie où ce type de données est très fréquemment rencontré. A l'issue de ce module, l'étudiant doit être capable de mettre en œuvre les méthodes enseignées pour résoudre les problèmes classiques de comparaison de plusieurs échantillons, de modélisations (paramétriques ou non) des distributions de survie.

Contenu de la matière

Les quatre premières séances seront consacrées à l'enseignement théorique des différentes méthodes. Une dernière séance consistera en un apprentissage des différentes procédures disponibles dans les logiciels (SAS en particulier). Le cours sera illustré de nombreux exemples et des exercices seront, à chaque étape, proposés aux étudiants pour vérifier qu'ils ont bien compris et acquis l'essentiel des méthodes. Ces exemples et exercices seront principalement issus de travaux effectués en recherche clinique et en épidémiologie. Cependant, le caractère général de ces méthodes, utiles dans bien d'autres domaines d'application, sera clairement établi.

1- Généralités et particularités des données de survie.
2-Définition des différentes fonctions de survie.
3-Recueil des données de survie et préparation de la base de données à analyser.
4-Estimations non paramétriques des fonctions de survie.
5-Estimations et comparaisons paramétriques des distributions de survie - Cas du modèle exponentiel. Modèle exponentiel généralisé. Etude graphique de l'adéquation de la modélisation.
6-Comparaisons non paramétriques de plusieurs distributions de survie : logrank pondérés, liens avec les tests de rangs.
7-Calcul du nombre de sujets nécessaires : cas du modèle exponentiel et du logrank.
8-Modélisation semi-paramétrique des fonctions de survie par le modèle de Cox : définition de la vraisemblance conditionnelle, problème de codages, modèles avec interactions, étude de l'adéquation de l'hypothèse des taux proportionnels.
9-Introduction à différents problèmes : variables dépendantes du temps, risques concurrents, recherche d'interactions qualitatives, analyses intermédiaires …
10-Etude d'un cas en vraie grandeur : Recherche de facteurs pronostiques chez des enfants atteints de tumeurs cérébrales.
11-Mise en application (Y. De Rycke).

Pré-requis

Aucun

Contrôle des connaissances

Examen écrit (2h et documents autorisés).

Références bibliographiques

- C. HILL, C. COM-NOUGUE, A. KRAMAR, T. MOREAU, J. O'QUIGLEY, R. SENOUSSI, C. CHASTANG, *Analyse statistique des données de survie (2e éd.),* Collection statistique en biologie et médecine, Flammarion, 1996.

Langue d'enseignement

Français

UE3 - Méthodologie statistique 2

Mesures de qualité de vie

Measuring Quality of Life

Cours : 12h • Atelier : 6h

Enseignants : Jean Benoît HARDOUIN (Université de Nantes)

Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

Les mesures de qualité de vie associée à la santé (Health-Related Quality of Life Measures) ont connu au cours des 20 dernières années un développement important, notamment dans le domaine de l'évaluation des stratégies thérapeutiques du cancer, du SIDA ou encore de maladies chroniques. Les mesures de qualité de vie font partie des « mesures subjectives en santé », par opposition aux mesures d'efficacité cliniques objectives, traditionnellement utilisées dans les évaluations. Elles font appel à des méthodes et concepts développés en psychométrie, mais aussi en économie de la santé (approche par les préférences individuelles).

Contenu de la matière

1. Introduction
   - 1.1. Contexte et approches : Les mesures de qualité de vie en santé
   - 1.2. Définition d'une échelle de qualité de vie – Typologie des échelles
2. L'approche psychométrique classique
   - 2.1. Construction d'un questionnaire (génération et réduction d'items, choix des dispositifs de réponse, méthodes de scoring des échelles multi-items)
   - 2.2. Les propriétés de fiabilité, validité, sensibilité au changement
     - 2.2.1. Fiabilité : mesures de reproductibilité et de cohérence interne (alpha de Cronbach)
     - 2.2.2. Validité structurelle et clinique des échelles
       - 2.2.2.1. Validation de la structure d'une échelle par analyses factorielles exploratoires (EFA) et confirmatoires (CFA – Modèles d'équations structurelles)
       - 2.2.2.2. Validation clinique d'une échelle : discrimination de groupes cliniques a priori
     - 2.2.3. Sensibilité au changement : indicateurs internes et externes (effect-sizes, courbes ROC …)
   - 2.3. Application : atelier de validation psychométrique d'une échelle de qualité de vie en cancérologie
3. L'approche économique des échelles de qualité de vie
   - 3.1. Rappel des méthodes d'évaluation économique en santé
   - 3.2. Les index de santé pondérés par les préférences des individus
4. Les nouvelles approches psychométriques : la Théorie de Réponse aux Items (Item Response Theory - IRT)
   - 4.1. L'IRT non paramétrique : Modèle de Guttman et modèles de Mokken
   - 4.2. Les modèles de la famille de Rasch
     - 4.2.1. Le modèle de Rasch
     - 4.2.2. Les modèles polytomiques de la famille de Rasch : Partial Credit Model et Rating Scale Model
   - 4.3. Les autres modèles paramétriques de la Théorie de Réponse aux Items
   - 4.4. Extensions : analyse longitudinale, modèles multidimensionnels, le fonctionnement différentiel des items, l'introduction de covariables
   - 4.5. Applications
     - 4.5.1. Exemple de validation d'échelle par un modèle IRT : le Knee injury and Osteoarthritis Outcome Score (KOOS)
     - 4.5.2. Exemple d'analyse de données par un modèle IRT : analyse spatio-temporelle de la qualité de vie en France
     - 4.5.3. Atelier machine sous SAS
     - 4.5.4.

Pré-requis

Contrôle des connaissances

Projet

Références bibliographiques

- NUNNALY, BERNSTEIN, *Psychometric Theory*, 3rd edition, 1994 Mc GRAW HILL
- P.M. FAYERS, D. MACHIN, Quality of Life : assessment, analysis and interpretation (2nd ed.), Wiley & Sons, 2007
- B. FALISSARD, *Mesurer la subjectivité en santé* (2e éd.) Collection évaluation et statistique, Masson, 2008
- K. Sijtsma, I. W. Molenaar. Introduction to Nonparametric Item Response Theory, Collection Measurement Methods for the Social Science. Sage Publications, Inc, 2002
- H. Fischer, I. W. Molenaar. Rasch Models: Foundations, Recent Developments, and Applications. Springer, 1995.

- Mesbah M, Cole BF, Ting Lee ML (editors) Statistical methods for Quality of Life studies : Designs, measurements and analysis. Luwer, 2002
- Streiner DL, Norman GR. *Health Mesaurement scales* (4th ed.). Oxford University Press, 2008

Langue d'enseignement
Français

UE3 - Méthodologie statistique 2

Statistique des processus

Statistics of Stochastic Processes

Cours : 12h Atelier : 3h

Enseignant : Myriam VIMOND **(**Ensai**)**

Correspondants : Salima EL KOLEI

Enseignement avec une partie commune aux élèves des filières « Statistique pour les sciences de la vie » et « Génie statistique »

Objectif pédagogique

Il s'agit de présenter des modélisations des principaux phénomènes aléatoires dépendants du temps rencontrés dans l'industrie et en sciences de la vie (hors-séries temporelles et traitement du signal) à partir de processus markoviens. Dans chaque cas la présentation portera autant sur les outils probabilistes que sur l'inférence statistique dans ces modèles.

Contenu de la matière

1. Compléments sur les chaînes de Markov. Théorèmes ergodiques. Estimation de la matrice de transition.
2. Processus de Poisson
3. Introduction aux chaînes de Markov cachés.

Pré-requis

Les cours de Probabilités et Statistique de première année et le cours de chaînes de Markov

Contrôle des connaissances

Examen écrit.

Références bibliographiques

- ASMUSSEN, S. Applied probability and queues. Second edition. Springer 2003.
- BOSQ, D. Statistique mathématique et statistique des processus. Lavoisier, 2012.
- DACUNHA-CASTELLE, D., DUFLO, M. Probabilités et statistiques II, Problèmes à temps mobile. Masson, 1993.
- DELMAS, J-F., JOURDAIN, B. Modèles aléatoires. Applications aux sciences de l'ingénieur et du vivant. Springer 2006.
- FOATA, D., FUCHS,A. Processus Stochastiques (2e éd.). Dunod 2004.
- FUCHS, C., Inference for Diffusion Processes, With Application in Life Sciences. Springer 2013.
- KIMMEL, M., AXELROD, D. Branching processes in biology. Springer 2002.
- PARDOUX, E. Processus de Markov et applications. Algorithmes, génome et finance. Dunod 2007.

Langue d'enseignement

Français

UE3 - Méthodologie statistique 2

Traitement des données manquantes dans les essais cliniques

Handling of missing data in clinical trials

Cours : 6h • Atelier : 6h

Enseignant : Mélanie PRAGUE (INRIA)

Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Contenu de la matière

PART I
Introduction to missing data – issues with missing data
Assumptions about missing data and missing data mechanisms – A formal taxonomy: MCAR, MAR and MNAR
Exercise 1
Regulatory considerations – The FDA-commissioned NRC/NAS report – Notion of estimand
Single imputation methods: LOCF, BOCF
Commonly used analytic methods under MAR:

- Complete-Case-Analysis (CCA)
- A modification of CCA under MAR: Inverse Probability Weighting (IPW)

PART II
Commonly used analytic methods under MAR (continued):

- Single imputation
- Multiple imputation
    - The multivariate normal regression framework
    - Monotone missing data patterns
    - Non-monotone missing data patterns: MCMC algorithms
    - Case of binary outcomes

PART III
Parametric analyses under MAR:

- Likelihood-based methods (Mixed Model with Repeated Measures = MMRM)
- Case study – Computer practical using SAS (version 9.4 or higher) – Questions 1-5
- Marginal versus conditional models
- Exercise 2

Introducing MNAR assumptions via the conditional model:

- Shift parameter
- Power considerations

Pattern-Mixture-Models (PMM)

PART IV
Principles and methods of sensitivity analyses under MNAR via PMM:

- Control-based imputation: Copy Reference (CR) and Jump To Reference (J2R) inference
- Case study – Computer practical using SAS – Question 6
- Delta adjustment and tipping point analysis
- Case study – Computer practical using SAS – Question 7
- Power considerations for MMRM and delta-adjusted PMM analyses
- Case study – Computer practical using SAS – Question 8

Key messages to take home

Contrôle des connaissances
A déterminer

Langue d'enseignement

Français mais supports écrits en anglais

UE3 - Méthodologie statistique 2

Statistique Bayésienne

Advanced Bayesian Statistics

Cours : 9h • Atelier : 6h

Enseignant : Sophie ANCELET (IRSN)

Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves de la filière "Génie Statistique", « Ingénierie statistique des territoires et de la santé » et « Statistique pour les sciences de la vie »

Objectif pédagogique

A l'issue de cet enseignement, les élèves devront maîtriser les connaissances de base pour l'analyse de données par approche statistique bayésienne. Les problèmes traités seront empreints aux sciences de la vie où l'emploi des méthodes bayésiennes progresse considérablement. Cependant, le caractère général de ces méthodes, utiles dans bien d'autres domaines d'application, sera clairement établi. À l'issue de ce module, l'étudiant doit être capable de mettre en œuvre les méthodes enseignées pour mener des inférences bayésiennes de données, notamment à l'aide des logiciels WINBUGS, OPENBUGS et JAGS.

Contenu de la matière

Un rappel de cours est fait concernant les principes de la modélisation statistique bayésienne. L'accent sera mis sur l'analyse bayésienne par les méthodes de Monte Carlo par Chaînes de Markov (MCMC). Aux travers d'exemples, seront abordés les notions de graphe d'indépendance conditionnelle, réseau bayésien, convergence des Chaînes de Markov, inférence, prédiction, validation et comparaison de modèles dans un cadre bayésien. Les exemples seront traîtés sous le logiciel WINBUGS ou JAGS en salle informatique.

Pré-requis

Cours de statistique bayésienne en deuxième année

Contrôle des connaissances

Projet court

Références bibliographiques

- Collectif BIOBAYES: Albert I., **Ancelet S.**, David O., Denis J.B., Makowski D., Parent E., Soubeyrand S**.** (2015) Méthodes statistiques bayésiennes. Bases théoriques et applications en alimentation, environnement et génétique. *ELLIPSES, ISBN : 978234000501*
- Carlin, B. P. and Louis, T.A. (2009). Bayesian Methods for Data Analysis. Chapman & HALL/CRC, third edition, (535 pp.)
- Gelman, A., Carlin, J. B., Stern, H. S. and. Rubin, D. B (2004). Bayesian data analysis. Texts in Statistical Science. Chapman & HALL/CRC, second edition, (668 pp.)
- Robert, C. P. (2001). The Bayesian choice. Springer, (second edition) (604 pp.)
- Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10: 325-337.
- Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). Markov chain Monte Carlo in practice. Chapman and Hall, (486 pp.)

Langue d'enseignement

Français

UE3 - Méthodologie statistique 2

Méta-Analyse

Meta Analysis

Cours : 12h • Atelier : 6h

Enseignant : Drifa BELHADI (Creativ Ceutical)

Correspondant : Samuel DANTHINE

Enseignement destiné aux élèves des filières « Ingénierie Statistique des Territoires et de la Santé» et « Statistique pour les sciences de la vie »

Objectif pédagogique

"La méta-analyse est une démarche, plus qu'une simple technique, qui a pour but de combiner les résultats de plusieurs essais thérapeutiques, pour en faire une synthèse reproductible et quantifiée. Cette synthèse produit un gain de puissance statistique dans la recherche de l'effet d'un traitement, une précision optimale dans l'estimation de la taille de l'effet et permet en cas de résultats apparemment discordants d'obtenir une vue globale de la situation".
Trois types de méta-analyses sont distingués, en fonction des données utilisées:

1. La méta-analyse des données résumées de la littérature, donc uniquement des essais publiés (ce qui expose au biais de publication)
2. La méta-analyse exhaustive sur données résumées se basant sur les études publiées et sur les travaux non publiés
3. La méta-analyse sur données individuelles se basant sur les données de tous les patients inclus dans les essais pris en considération dans la méta-analyse.

Dans la démarche de la méta-analyse, la variabilité (l'hétérogénéité) est considérée comme un paramètre de nuisance; elle contredit l'hypothèse de l'existence d'un effet traitement commun à tous les essais. La méta-analyse est très utilisée, notamment dans les analyses médico-économiques qui utilisent dans leur modélisation des indicateurs de résultats issus de publications diverses.

Contenu de la matière

1. Introduction
   - 1.1. What is it? / Why do we do it?
   - 1.2. The use of meta-analysis in clinical trial / Health economic evaluation
   - 1.3. Meta-analysis vs. randomised clinical trials
2. Protocol development
   - 2.1. Objectives
   - 2.2. Outcome measure and baseline information
   - 2.3. Data sources / Study selection
   - 2.4. Data extraction
   - 2.5. Analyses / Sensitivity analyses
   - 2.6. Presentation of results
3. Estimating treatment difference
   - 3.1. Binary data
     - 3.1.1. Log-odds ratio
     - 3.1.2. Log-relative risk
   - 3.2. Normally distributed data
     - 3.2.1. Absolute mean difference
     - 3.2.2. Standardised mean difference
   - 3.3. Ordinal data
     - 3.3.1. Log-odds ratio (prortional odds model)
   - 3.4. Survival data
     - 3.4.1. Log hazard ratio
4. Combining estimates of treatment difference
   - 4.1. Fixed-effects parametric approach (FE)
     - 4.1.1. Definition/assumption
     - 4.1.2. Model
     - 4.1.3. Estimation of the treatment difference and hypothesis test
     - 4.1.4. Testing for heterogeneity
   - 4.2. Random-effect parametric approach (RE)
     - 4.2.1. Definition/assumption
     - 4.2.2. Model
     - 4.2.3. Estimation of the treatment difference and hypothesis test
     - 4.2.4. Testing for between studies heterogeneity
5. Dealing with heterogeneity
   - 5.1. Limited power of heterogeneity tests
   - 5.2. Choice between FE and RE models
   - 5.3. Can we always present an overall estimate of treatment difference?
   - 5.4. Choice of appropriate measure of treatment difference
   - 5.5. Meta regression

6. Presentation of results
7. Selection / publication bias
8. Direct comparison vs. Indirect comparison
   8.1. Eg. Drug A vs placebo & Drub B vs. placebo => Drug A vs. Drug B
9. An introduction to Bayesian approach
10. Conclusion
    10.1. The use of meta-analysis
    10.2. Contrast between useful and useless meta-analysis

Pré-requis

Basic Winbugs knowledge

Contrôle des connaissances
A déterminer

Références bibliographiques
Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Chichester (UK): John Wiley & Sons, 2008.

Dias, S., Welton, N.J., Sutton, A.J. & Ades, A.E. NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials. 2011; last updated September 2016; available from http://www.nicedsu.org.uk

Langue d'enseignement
Anglais

UE4 - Essais cliniques

Essais cliniques : méthodologie et analyses statistiques

Clinical Trials

Cours : 18h

Enseignants : David HAJAGE (INSERM APHP) Responsable du cours, Lisa BELIN et Yann DE RYCKE

Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves des filières « Statistique pour les sciences de la vie » et « Ingénierie des territoires et de la santé »

Objectif pédagogique

La nature et la structure des données recueillies dans le cadre d'essais cliniques (qui incluent les études sur les médicaments, les interventions médicales novatrices et les nouveaux matériels) nécessitent de recourir à des méthodes statistiques adaptées.
Cet enseignement permettra aux élèves de se familiariser avec les différents types d'études, les enjeux, les acteurs et plus particulièrement les méthodes statistiques utilisées dans le domaine des essais cliniques.

Contenu de la matière

Après une présentation générale des essais cliniques, le cours comportera 2 parties. La première aura comme objectif de permettre aux élèves de se familiariser avec la méthodologie des essais cliniques et de découvrir le déroulement d'une étude du point de vue du biostatisticien. La seconde s'attachera à détailler certaines méthodes utilisées dans l'analyse des études cliniques.

Présentation générale des essais cliniques :
Les différentes étapes d'une étude, les intervenants, le rôle du biostatisticien
Aspects règlementaires et éthiques
Déroulement d'une étude pour le biostatisticien
L'analyse statistique : du plan d'analyse aux résultats
Choix d'une méthode adaptée aux objectifs et aux données
Puissance et nombre de sujets nécessaires
Rédaction d'un rapport statistique
Divers éléments à prendre en considération : Biais, Indépendance, Normalité, Bilatéral / Unilatéral, Populations ITT et PP, …
Approfondissement de quelques méthodes d'analyse
Mesures répétées
Essais de différence, d'équivalence, de supériorité, de non infériorité
Courbes ROC
Concordance, fiabilité, reproductibilité

Contrôle des connaissances

Examen écrit

Références bibliographiques

Sera distribuée en séance

Langue d'enseignement

Français

UE4 - Essais cliniques

Pharmacométrie

Pharmacometrics

Cours : 6h • Atelier : 12h

Enseignants	:	Jérémie GUEDJ, Emmanuelle COMETS (INSERM, Institut Claude Bernard University Paris Diderot)
Correspondant	:	Lionel TRUQUET

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique
A l'issue de cet enseignement, les élèves devront maîtriser les principales méthodes statistiques utilisées par le biostatisticien lors de l'analyse ou la conception des essais cliniques.

Contenu de la matière
Introduction à la pharmacométrie, la pharmacocinétique et la pharmacodynamie (principes, rôle dans le développement des médicaments, exemples de modèles)
Modèles non-linéaires à effets mixtes (introduction, historique, modèles statistiques, méthodes d'estimation, logiciels)
Construction et évaluation de modèles
TP en Monolix (présentation du logiciel, construction d'un modèle pharmacocinétique)

Optimisation de protocoles dans les modèles non linéaires simples ou mixtes, théorie et applications, simulation d'essais cliniques
TP en Monolix et PFIM (graphes diagnostiques, simulation, optimiation)
Séminaire

Pré-requis
Modèles Mixtes

Contrôle des connaissances
Examen

Références bibliographiques

Langue d'enseignement
Français

UE4 - Essais cliniques

Projet d'essais cliniques

Project in Clinical Trials

Atelier : 6h • Projet : 18h

Enseignant : David HAJAGE (INSERM APHP)

Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

Le but de ces projets est de mettre en application quelques-unes des méthodes vues pendant le cours sur les essais cliniques. Dans cet objectif, une base de données correspondant à un essai clinique réel est fournie aux élèves, ainsi que les documents nécessaires (cahier d'observation, protocole...) afin qu'ils réalisent suivant des directives précises l'analyse demandée, en utilisant le logiciel R.

Contenu de la matière

Ce projet, réalisé en général par groupe de 2 ou 3 élèves, demande un important travail de programmation, de réflexion statistique, un effort de recherche sur la pathologie étudiée ainsi qu'une bonne compréhension des recommandations internationales sur l'analyse statistique et la rédaction d'un rapport dans le cadre d'un essai clinique. A la fin du projet, les élèves doivent remettre un rapport statistique d'une trentaine de pages donnant lieu à une soutenance réalisée devant des professionnels des essais cliniques.
Le rapport ainsi que la soutenance seront réalisés en anglais, langue de travail utilisée dans l'industrie pharmaceutique. Les élèves auront à ce titre un soutien assuré par un enseignant d'anglais.

Pré-requis

Introduction aux essais cliniques, compléments d'essais cliniques. Bonne maîtrise des méthodes statistiques générales, en particulier: statistiques descriptives, régression linéaire, analyse de variance, régression logistique, analyse des données de survie, modèles mixtes.

Contrôle des connaissances

L'évaluation des projets repose à la fois sur le contenu des rapports et le comportement à la soutenance, devant un jury composé de spécialistes.

Références bibliographiques

Sera distribuée en séance

Langue d'enseignement

Français. Le rapport sera rédigé en anglais, et donnera lieu à une soutenance en anglais.

UE5 - Epidémiologie

Epidémiologie quantitative

Quantitative Epidemiology

Cours : 18h

Enseignant : Olivier GRIMAUD , Pascal CREPEY, Cindy PADILLA (EHESP)

Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

L'objet de l'épidémiologie est l'étude la distribution dans le temps et dans l'espace des états de santé des populations humaines et l'analyse leurs déterminants.

A l'issue du cours, l'étudiant sera capable de :

- Procéder à une description simple de la situation épidémiologique à partir des indicateurs épidémiologiques usuels.
- Formuler les hypothèses de liaison, discuter de la possibilité d'une relation de cause à effets entre les facteurs de risque et l'état de santé.
- Choisir les types d'études appropriés à la mise en évidence des relations entre maladies et facteurs de risque.
- Interpréter les résultats d'études épidémiologiques.

Contenu de la matière

Intitulé de la séance de cours
Introduction de l'épidémiologie – Causalité – Enquêtes de cohorte
Epidémiologie descriptive: prévalence, incidence, standardisation
Les mesures d'association
Enquêtes cas-témoins. Biais
Facteurs de confusion
Analyses multivariées

Pré-requis

Les méthodes statistiques de base sont supposées connues (fluctuations d'échantillonnage, intervalle de confiance, méthodes d'estimation et tests statistiques usuels).

Contrôle des connaissances

L'évaluation consistera en un examen écrit de 1 heure et demi, tous documents et calculette sont autorisés.

Références bibliographiques

- J. BOUYER, D. HEMON, S. CORDIER, F. DERRIENNIC, I. STUCKER, B. STENGEL, J. CLAVEL, *Epidémiologie. Principes et méthodes quantitatives,* Paris, Les éditions INSERM, 1993
- K.J. ROTHMAN, S. GREENLAND, *Modern epidemiology,* Little, Brown and Company, Boston, 1998
- D.G. KLEI NBAUM, L.L. KUPPER, H. MORGENSTERN*, Epidemiologic Research. Principles and quantitative methods,* New York, Van Nostrand Reinhold, 1982

Langue d'enseignement

Français

UE5 – Epidémiologie

Modélisation compartimentale

Compartmental Modeling

Cours : 6h • Atelier : 6h

Enseignant	Audrey LAVENU (Université de Rennes 1)
Correspondant	Lionel TRUQUET

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique
A l'issue de cet enseignement, les élèves devront être capables de simuler des épidémies par des modèles Susceptible-Infectious-Removed (SIR), de comprendre l'interprétation des paramètres et de construire des modèles dérivés du modèle standard.

Contenu de la matière
Introduction : Contexte de la modélisation compartimental. Définition d'une épidémie.
1. Modèle SIR déterministe
1.1. Construction du modèle (système d'équations différentielles)
1.2. Calcul du taux de reproduction de base (R0)
1.3. Simulation d'épidémies
1.4. Exemples de données de surveillance épidémiologique
- le réseau Sentinelles
- application sur logiciel

1.5. Exemples de modèles dérivés du modèle SIR.
2. Modèle SIR stochastique

Pré-requis
Aucun

Contrôle des connaissances
Rapport basé sur exercices commencés en séance avec commentaires étoffés du cours

Références bibliographiques

- Anderson, R. A. and May R. M. 1982. Directly transmitted infectious diseases: Control by vaccination, Science 215, 1053-1060.
- Anderson R. A. and May R. M., 1992: Infectious Diseases of Humans: Dynamics and Control (2nd ed.). Oxford University Press, Oxford.
- Bailey N. T. J., 1975: The mathematical theory of infectious diseases and its application. Griffin, London, 2nd edition (épuisé).
- Bartlett M. S. 1960. Stochastic Population Models in Ecology and Epidemiology,
- Methuen, London.

Langue d'enseignement
Français

UE6 - Statistique pour données Omics

Analyse des données « Omics »

Omics Data Analysis

Cours : 6h • Atelier : 12h

Enseignants : Isabelle BRITO (Institut Curie) et Pierre NEUVIAL (IMT)

Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique

Ces séminaires permettent aux étudiants de rencontrer différents chercheurs travaillant sur l'analyse d'une grande variété de données « omics », y compris les données génomiques, transcriptomiques, métabolomiques, protéomiques, épigénétiques, et métagénomiques.

Contenu de la matière

Les technologies telles que le séquençage à haut débit, les puces à ADN,et la spectrométrie de masse ont modifié l'échelle de données biologiques disponibles et permettent de générer des quantités de données importantes à plusieurs niveaux biologiques. L'analyse de ces données fait appel à de nombreuses méthodes statistiques qui seront présentées.

Plusieurs problèmes statistiques se posent dans l'analyse des données transcriptomiques : normalisation des données, modélisation des variances pour la construction des statistiques de tests afin de détecter les gènes différentiellement exprimés, corrections pour les tests multiples, classification et modèles de mélange afin d'obtenir des groupes de gènes à fonction biologique similaire. Des approches de la biologie des systèmes seront également abordées, tels que l'inférence de réseaux de gènes à partir de données transcriptomiques. Les questions posées par l'analyse de données épigénétiques et métagénomiques seront brièvement abordées. Le cours est accompagné de travaux pratiques (sur R/Bioconductor) centrés sur une analyse différentielle de données issues des puces à ADN et du séquençage à haut débit (RNA-seq). Des problématiques statistiques similaires seront abordées pour l'analyse de données métabolomiques.

Pré-requis

Introduction à l'analyse de données « Omics », théorie des tests, modèle linéaire (ANOVA et modèle mixte), classification.

Contrôle des connaissances

Un examen écrit qui comprend les notions abordées en cours et des extraits de scripts R.

Références bibliographiques

- Mary-Huard T., Picard F., Robin S. Introduction to Statistical Methods for Microarray Data Analysis, in *Mathematical and Computational Methods in Biology*, Hermann : Paris, 2007.
- Gentleman R.C., Carey V.J., Dudoit S., Irizarry R., Huber W., *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, New York: Springer, 397-420, 2005.

Langue d'enseignement

Français

UE6 - Statistique pour données Omics

Introduction à l'analyse de données « Omics »

Introduction to Moics Data Analysis

Cours : 6h • Atelier : 6h

Enseignant : Guillaume DESACHY (IRIS Servier)

Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves de la filière « Statistique pour les sciences de la vie »

Objectif pédagogique
Les séminaires sur le génome permettent aux étudiants de rencontrer différents chercheurs travaillant sur des problèmes liés aux génomes.

Contenu de la matière
L'étude de la génomique a de nombreuses applications dans le domaine des sciences de la vie. Les problèmes liés au génome font appel à de nombreuses méthodes statistiques. Différentes approches pour l'analyse des séquences génomiques seront présentées.
Le cours consiste en une vaste initiation à la Génomique et Post-Génomique (Transcriptomique, Protéomique) et de façon plus générale à la Bioinformatique. Les objectifs de ce cours sont l'acquisition des connaissances de bases (du Génome à l'Organisme en passant par l'ADN, l'ARN, la protéine, la cellule et les bases de données associées) permettant la compréhension des principales problématiques liées à l'application de la Statistique à la Biologie. Les principales méthodes statistiques évoquées sont les chaînes-de-markov, chaînes-de-markov cachées, test d'hypothèse, test-multiple, la classification et l'alignement de séquences biologiques. Le cours est accompagné de travaux pratiques réalisés sous *R*.

Pré-requis
Pas de pré-requis

Contrôle des connaissances
Mini-projets en anglais, compte-rendus de TP.

Langue d'enseignement
Français

UE – Projet de fin d'études

Projet méthodologique

Methodological project

Atelier : 3h • Projet : 9h

Enseignants : Divers intervenants

Correspondant : Arthur KATOSSKY (Ensai)

Enseignement destiné à l'ensemble des élèves des six filières

Objectif pédagogique

Le projet méthodologique consiste en la production d'un article de synthèse sur un sujet de recherche à choisir parmi un catalogue. Ses objectifs sont multiples:

- familiarisation avec la forme des productions académiques
- mise en œuvre d'une démarche scientifique rigoureuse
- prise de conscience des enjeux autour de la reproductibilité des résultats de recherche
- travail en binôme
- communication sur des sujets techniques

À cela s'ajoute les objectifs spécifiques à la production d'un travail scientifique en langue anglaise (expression écrite et orale, vocabulaire spécialisé, vulgarisation, etc.).

Contenu de la matière

Travail de recherche en groupe suivi par un chercheur (env. 5 séances) et un professeur d'anglais (4 séances).

Contrôle des connaissances/ Evaluation

Projet

Références bibliographiques

Selon les projets

Langue d'enseignement

Anglais

UE – Projet de fin d'études

Projet de fin d'études

End of study project

Atelie : 9h - Projet : 27h

Enseignant	:	Divers intervenants industriels
Correspondant	:	Arthur KATOSSKY (Ensai

Enseignement destiné à l'ensemble des élèves des six filières

Objectif pédagogique
Le projet de fin d'études consiste en la production d'une étude statistique de niveau professionnel dans le monde de l'entreprise ou de la recherche, parmi un catalogue de sujet mis à disposition des élèves. Ses objectifs sont multiples:
- mise en situation professionnelle
- capacité à définir une stratégie d'étude en réponse à une demande client
- mobilisation des compétences techniques (statistiques, économiques, informatiques)
- compromis entre rigueur scientifique et contraintes pratiques (limitations financières, logicielles, cognitives, temporelles…)

travail de groupe
- gestion d'un projet sur le temps long
- communication (écrite, orale) sur des sujets techniques

Contenu de la matière
Travail autonome en groupe suivi par un professionnel de l'entreprise ou de la recherche (env. 5 séances)

Pré-requis

Références bibliographiques
Selon les projets

Contrôle des connaissances
Évaluation: projet avec soutenance

Langue d'enseignement
Français

UE – Projet de fin d'études

Data challenge

Data Challenge

Atelier : 12h

Enseignant : Divers intervenants industriels

Correspondant : Salima EL KOLEI

Enseignement destiné à l'ensemble des élèves des six filières

Objectif du cours

Le data challenge permet de rassembler sur une période très courte différentes équipes de profils variés afin de collaborer sur un projet. Cette expérience se rapproche des conditions réelles dans laquelle évoluent les datascientists au sein des entreprises.
Il permet, à partir des mécanismes du jeu, de dynamiser et d'articuler la pédagogie autour d'un besoin concret d'entreprise et d'un événement qui s'achève par une évaluation objective. De nombreux challenges sont proposés autour de la Data ou présentant des problématiques Data importantes.
L'objectif de ce cours est de valoriser et d'évaluer les compétences transversales acquises dans ce contexte opérationnel.

Contenu de la matière

Les élèves devront participer au data challenge proposé à l'Ensai ouvert également aux élèves de deuxième année.

Compétences acquises

-Comprendre les problèmes à résoudre.
-Travailler en mode projet avec des contraintes.
-S'intégrer et s'adapter dans un contexte pluridisciplinaire. Selon les challenges, les compétences seront mobilisées à géométrie variable.
-S'adapter à la réalité de la Data d'entreprise (données non structurées, manquantes, volumétrie…)
-Communication orale des résultats (pitch…)

Pré-requis

-Compétences en statistiques et informatiques de 1A, 2A et 3A.
-Compétences transversales mobilisées dans les projets 1A, 2A et 3A.

UE - Séminaires professionnels

Evaluation médico-économique

Economic Evaluation in Health Care

Atelier : 12h

Enseignant : Hélène CAWSTON et Lueza BERANGER (Amaris)

Correspondant : Samuel DANTHINE

Enseignement destiné aux élèves des filières « Statistique pour les sciences de la vie » et « Ingénierie des territoires et de la santé »

Objectif pédagogique

L'objectif de ce cours est de présenter les méthodes de l'évaluation économique dans le domaine de la santé. Ce cours abordera d'abord les principes fondamentaux de l'évaluation sur le terrain, puis les différentes méthodes seront présentées et discutées, notamment du point de vue de résultats utilisés : l'évaluation coût-efficacité, l'évaluation coût-utilité, l'évaluation coût-bénéfice. Les aspects théoriques et techniques de l'évaluation médico-économique seront abordés, en particulier les techniques de modélisation statistique avancés permettant d'intégrer l'incertitude dans le calcul économique.

Contenu de la matière

1. Introduction à l'évaluation économique en santé
   - 1.1. Concepts
   - 1.2. Méthodes d'évaluation
2. Modélisation économique
   - 2.1. Arbres de décision
   - 2.2. Modèles de MARKOV
3. Analyse de l'incertitude
   - 3.1. Analyse classique
   - 3.2. Analyse probabiliste
4. Analyse critique d'un article

Pré-requis

Contrôle des connaissances

Examen écrit

Références bibliographiques

- M.F. DRUMMOND, B.J. O'BRIEN, G.L. STODDART, G.W. TORRANCE, *Méthodes d'évaluation économique des programmes de santé*, Economica, 2ème édition, 1998
- M.F. DRUMMOND, McGUIRE A, Economic evaluation in health care, 2001.

Langue d'enseignement

Français

UE - Séminaires professionnels

BIG DATA, data Mining et machine Learning

BIG DATA, data Mining and machine Learning

Atelier : 12h

Enseignant : Xavier VAN AUSLOOS (LA DONNEE INTELLIGENTE)

Correspondant : Lionel TRUQUET

Enseignement destiné aux élèves de la filière «Statistique pour les sciences de la vie»

Objectif pédagogique

Introduction aux problématiques BIG DATA, data Mining et machine Learning en santé

Contenu de la matière

- Métier et les enjeux en santé :
  - projets d'innovations avec Institut Paoli Calmettes de Marseille autour de la médecine de précision/traitement
  - projet d'innovation mené par une collègue de SOGETI HT : http://open-intelligence.fr/genomique-et-big-data-lassociation-integragen-igr-inserm-et-sogeti-hightech/
  - projet SIRIC : http://www.e-cancer.fr/Professionnels-de-la-recherche/Recherche-translationnelle/Les-SIRIC: mise en place d'une infra type Big Data pour la recherche translationnelle en cancérologie
  - freins actuels : sécurité, peu de connaissances techniques, confidentialité des données
  - projet HANDILIGHT http://www.lemondeinformatique.fr/defih/zone-presse.html
- Intro au Big Data :
  - pourquoi le big data vs info traditionnelle
  - Ecosystèmes
  - Hadoop / HDFS
  - Hive/Spark
  - noSQL
  - Data Processing with YARN/MapReduce2

- TD etTP
  - Lab 1 Hadoop Hands-On Lab
  - Hadoop Languages
  - Lab 2 Hadoop Languages
  - Hive
  - Lab 3 Hive
  - Spark Fundamentals
  - Lab 4 Spark Fundamentals
  - Lab 4 Spark Fundamentals (continued)
  - Using SQL with Hadoop (IBM BigSQL or equivalent)
  - Using R with Hadoop (IBM BigR and equivalent)
  - HBase
  - Lab 5 HBase
  - Ambari
  - Lab 6 Ambari
  - IBM BigSheets, MS PowerBI,
  - Lab 7 BigSheets or MS PowerBI
  - Text Analytics Fundamentals
  - Lab 8 Text Analytics Fundamentals lab
  - Final Considerations

Pré-requis

Contrôle des connaissances

Pas d'évaluation.

Références bibliographiques

UE – Séminaires professionnels

Statistique des processus

Statistics of Stochastic Processes

Atelier : 6h

Enseignant : Maud Delattre **(Agrocampus**)

Correspondants : Lionel Truquet

Objectif pédagogique
Il s'agit de présenter des exemples d'utilisation de chaînes de Markov en oncologie et en épidémiologie.

Pré-requis

Le cours de statistique des processus

Contrôle des connaissances
Pas d'évaluation.

Langue d'enseignement
Français